An Imputation Approach to the Fusion of Travel Diary and Lifestyle Data: Application to the Analysis of the Interaction of ICT and Physical Mobility

Jacek Pawlak[1], John Polak[2] and Aruna Sivakumar[3]
[1]Imperial College London, jacek.pawlak09@imperial.ac.uk
[2]Imperial College London, j.polak@imperial.ac.uk
[3]Imperial College London, a.sivakumar@imperial.ac.uk

Abstract

In the light of recent and on-going developments in information and communication technologies (ICT), there is an ever growing interest in the relationship between digital behaviour and physical mobility. However, data shortages together with privacy and ethical concerns can limit comprehensiveness of the empirical studies dealing with the issue. Data fusion (pooling), whereby different datasets are combined into a single synthetic source and dealt with appropriately to draw approximately valid inferences, offers a promising solution to this problem. In this paper, we present an implicit method of data pooling using the k-Nearest Neighbour algorithm for combining information on individuals' digital lifestyles and travel behaviour. We employ this method to datasets from 3 countries: the US, UK, and Norway, and using multiple imputations approach derive approximately valid inferences from the developed structural equation models (SEM). We also suggest potential improvements to the technique which remain the subject of our continuing research.

Keywords: data pooling, k-NN algorithm, structural equation modelling

1. Introduction

Recent and on-going developments in technology and policy are increasingly disrupting traditional socio-economic boundaries and behaviours. For example, the growth of e-commerce and m-commerce is radically changing the nature of retail activity, with consequential impacts on individual travel behaviour, retail logistics, work practices and the locational decisions of individuals and businesses. Similarly, climate change-related policies aimed at de-carbonising urban mobility (e.g., the promotion of electric mobility) are blurring the boundary between consumer behaviour in the transport and energy sectors. These developments give rise to many new challenges for researchers and policy makers, one of which is the need to understand behaviour more holistically, often cutting across the demarcations of traditional survey instruments. Although it is sometimes possible to create new survey instruments to address these specific new requirements, significant administrative and financial constraints are usually present limiting this approach in practice. Therefore, increased attention is being focused on novel methods of combining existing official statistical sources in ways that address these new challenges.

In this paper we describe a method of combining common types of official data sources – travel diary and time use based mobility data and lifestyle (Internet use) data. The method

operates at the unit record level to produce a fused synthetic dataset, in which each record comprises information from each dataset. The paper presents the statistical methods involved and describes an application to the study which combines travel diary and Internet use data to study interactions between digital behaviour and physical mobility. As such, it is organised into a number of sections. Section 2 presents the motivation and context of the work. Section 3 presents the proposed data fusion methodology and describes its practical implementation in the context combining mobility and lifestyle data, using datasets from the US, UK and Norway. Section 4 presents the application of the fused datasets to the analysis of the interaction of digital lifestyles and physical mobility using a structural equation modelling (SEM), discusses the usefulness of the method, as well as the associated challenges and possible improvements. The final section presents some overall conclusions from the work.

2. Motivation and Context

The motivation for the development of the method described in this paper is the growing need to better understand the relationship between digital behaviour and physical mobility. The need to understand, explain and model these relationships has been an important research topic in the domain of transport studies since at least 1980s (Salomon 1986; Mokhtarian 1990). However, in recent years the growth both in the range and penetration of ICT-based services and activities (e.g., e-leisure, e-entertainment, e-banking, e-education, e-government, e-healthcare, e-commerce etc.) has led to renewed interest (Golob and Regan 2001). Whilst some researchers have argued that the existence of increasing opportunities to undertake activities in a virtual mode will lead to a reduction in physical mobility, the empirical evidence suggests considerably greater complexity (Salomon 2000; Andreev et al. 2010).

The impacts of ICT on travel behaviour are rarely confined to reduction in travel demand. On the contrary, they typically involve changes in the whole activity pattern, such as travel timing, frequency, destination, purpose, mode, sharing patterns, and also increase in travel demand (Senbil 2009; Singh et al. 2011). Indeed, even the assumption of clear-cut distinction between physical and virtual activities has started to become blurred, especially in the presence of simultaneous activities, e.g., working on laptop while travelling on a train (Lyons and Urry 2005; Pawlak et al. 2011). A number of researchers studied these complex relationships using a flexible and convenient SEM approach (Golob 2003; Choo and Mokhtarian 2007; Wang and Law 2007). However, untangling the effects of technological developments and changes in digital lifestyles on travel behaviour and providing credible explanations and predictions requires simultaneous unit level detailed information on digital activity and lifestyle, physical mobility and relevant background demographic factors.

No single data source can provide all these data. On the one hand, conventional travel diary data of the type collected by many European countries records unit level physical mobility and certain demographics but contains no information on digital lifestyles (Stopher 1992, 2000; Axhausen 1998, 2008). On the other hand, Internet use, lifestyle data or opinions surveys contain richer and richer information on digital activity and demographics but only the most rudimentary information, if at all, on mobility (Hamermesh et al. 2005; Noce and McKeown 2008). Additionally, more and more sophisticated methods of passive data collection, e.g., via GPS and GSM enabled devices, or specialised transaction monitoring software raise significant privacy and ethical concerns regarding the extent to which a single dataset can simultaneously include information on different aspects of an individual's activity. However, increasingly the application to advanced models requires rich (multi-

sectoral) micro data, hence clashing with such concerns. Given the circumstances, the pooling of data from separate survey to generate a complete, but synthetic dataset, would appear to be an attractive ways of addressing these challenges. The importance of data pooling along with a number of examples of its applicability is discussed in more detail in Sivakumar and Polak (2013).

3. Data Fusion Methodology

In this study we adopt an imputation approach in which the data fusion (pooling, or grafting as termed by Saporta 2002) problem is framed as a missing data problem (Aluja-Banet et al. 2007). In particular, we treat the lifestyle (digital behaviour) data as having missing information on physical mobility. On the other hand, the second dataset (travel data), include detailed information on physical mobility which can be utilised to impute the missing variables in the former, digital behaviour dataset. Thus the lifestyle dataset is a target or recipient dataset while travel a donor or training dataset (D'Ambrosio et al. 2007). What is crucial, however, is that both datasets share a number of variables (shared, or common variables) which describe number, usually socio-demographic, features of the respondents, e.g. household car ownership. At this stage one faces the choice between parametric, explicit and implicit technique of imputation. The parametric and explicit methods develop, using shared and to-be-imputed variables, statistical models, e.g. linear regression, or estimates suitable parameters for the posterior Bayesian distribution of the missing information. The implicit technique is based on the idea of measuring similarity between each respondent in the target dataset and all respondents in the training dataset, with the most similar respondent in the training dataset donating the missing variables to the respondent in the target dataset (Saporta 2002; Aluja-Banet et al. 2007; Sivakumar and Polak 2013).

The main advantage of the explicit approach is in its capability of measuring the robustness using various goodness-of-fit statistics, e.g. coefficient of determination. In following such an approach a priori assumptions must be made, be it regarding the type of the regression fit or a priori Bayesian distribution. However, certain assumptions, such as linear regression, can induce various undesirable statistical phenomena in subsequent analysis of the synthesised complete dataset. Moreover, an accurate imputation of *a set* of variables simultaneously, such as daily travel durations for different travel purposes, requires development of a model capable of capturing the underlying correlation structure (Saporta 2002). Otherwise, one risks between-variable inconsistency, e.g. daily travel durations summing up to 25 hours. In such a case, structural equation modelling appears tempting, but largely linear in nature (despite existent of more complex non-linear approaches) the approach coulde possibly introduce multicollinearity and compromise the subsequent analysis of the full, synthesised dataset.

Consequently, in the context of our research we decided to follow an alternative, implicit path. In this case, the missing mobility information is still donated by the travel diary data, but the donation is based on matching similar records in the donor (travel) and recipient (lifestyle) datasets. The matching is performed using a k-Nearest Neighbour (k-NN) algorithm that operates on the basis of the shared (common) variables. Being non-parametric, the method relies to a lesser extent on making potentially erroneous or unverifiable distributional assumptions. Secondly, since the whole set of variables is transferred from the donor, the consistency between the imputed variables is ensured (Aluja-Banet et al. 2007). Thirdly, the method is flexible in terms both of its treatment of the data and the metrics of similarity that can be used. Finally, it naturally generates multiple candidate donor records,

thereby enabling the application of a multiple imputations framework (Rubin 1987) reducing the biases potentially affecting subsequent inferences (as discussed in the next paragraph). However, the non-parametric nature of the method makes quantification of its robustness challenging, and largely subject of the on-going research, including ours (see section 4 for more details). Moreover, it does not deal with the problem of the missing data present in the training data set due to the actual non-response. While respondents with the missing values can be excluded from the procedure, this could significantly reduce the training set, and thus conventional methods of non-response treatment are usually applied. Still, according to Saporta (2002, p. 471), implicit fusion is "efficient in keeping covariance structure and avoiding incoherencies" which is highly desirable for a subsequent estimation SEM which relies on maximising the fit between observed and modelled covariance matrices.

Regardless of the approach, imputing only a single value, i.e. creating only one fused dataset, carries a number of significant risks (Rubin 1987). First of all, it assumes that the imputation framework can perfectly predict the desired value, thus neglecting any stochastic variation in the imputed variables. Secondly, such a naïve approach underestimates the variability in the data, leading to underestimation of the standard errors in any subsequent derived parameter estimates. This effectively increases risk of type 1 error (incorrect rejection of a true null hypothesis, i.e. not rejecting lack of association between digital lifestyle and travel behaviour). In order to mitigate such problems, one introduces a degree of randomness in imputation. In the explicit case, this is achieved by adding a randomly distributed error to the inputed data, while in for the implicit approach, a set of $\boldsymbol{k}$ most similar respondents (nearest neighbours) are identified. A Monte Carlo draw from this set provides values for the target dataset with the probability of selection dependent on the similarity to the target respondent. Having in such a manner obtained $\boldsymbol{m}$ complete datasets, one can under certain conditions (described in detail by Rubin 1987) draw valid inferences about the estimates in the final analysis. This is done by using estimates of the desired parameters μ_n obtained in each imputation to obtain between-imputation-variance σ^2 (1) and combine it with the mean variance of the estimates s^2 (2):

$$\sigma_l^{\,2} = \frac{1}{m}\sum_{n=1}^{m}(\mu_n - \bar{\mu})^2 \tag{1}$$

$$\hat{\sigma}_l^{\,2} = \bar{s}_l^2 + (1 + m^{-1})\sigma_l^{\,2} \tag{2}$$

Thus (2) is an expression for the final variance of the estimate, and takes into account the stability of the final estimates of the parameters by including the between-imputations variability, and is also dependent on the number of imputations $\boldsymbol{m}$. In doing so, the method accounts for the reduced information content due to the data missingness. A caveat should be made here, that (2) has been proven by Rubin to be appropriate in the context of proper Bayesian models while it may underestimate the variability of the data for implicit, metric-matching imputations. However, his interpretation was in the context of survey non-response where the values are drawn from and imputed to the same dataset. However, the concern is less where the values are drawn from a complete sample with an unaffected, although unknown, distribution of the to-be imputed values.

An important aspect of the k-NN matching procedure is the choice of the distance metric, or (dis)similarity measure between the respondents. We have chosen to most widely applied measure, i.e. the Mahalanobis distance:

$$D(\vec{x_i}, \vec{x_j}) = \sqrt{(\vec{x_i} - \vec{x_j})^T \Sigma^{-1} (\vec{x_i} - \vec{x_j})} \tag{3}$$

where $\boldsymbol{x}$ is a vector of $\boldsymbol{n}$ comparable metrics characterising respondents $\boldsymbol{i}$ and $\boldsymbol{j}$, and $\boldsymbol{\Sigma}$ is a covariance matrix. Clearly, if $\boldsymbol{\Sigma}$ is an identity matrix, Mahalanobis distance is equivalent to Euclidian distance in $\boldsymbol{n}$ dimensions. The inclusion of covariance matrix makes the final measure independent of the units of comparable metrics (shared variables). Additionally, the possible correlation between these variables is taken into account and discounted so that variables representing correlated features of an individual, e.g. income and expenditure, would be weighed down appropriately ensuring more effective comparison. Calculation of the Mahalanobis distance is pretty straightforward in case of continuous variables. However, it becomes more challenging in the presence of discrete ordinal and nominal, variables. In such a case, not only the question is of how to represent numerically the difference between categories, e.g. genders, but also how to appropriately capture the underlying correlation structure. In fact, fairly complex modifications of Mahalanobis distance have been devised to enable inclusion of mixed data (see, for instance, Leon and Carrière 2005).

We propose a simpler, alternative solution to inclusion of ordinal and nominal variables in Mahalanobis distance based on the idea of symbolic covariance (McCane and Albert 2008). The core concept, in this case, is a function $\boldsymbol{\delta(x_i, x_j)}$ attaching numerical values to differences in discrete categories of the variable $\boldsymbol{x}$ characterising respondents $\boldsymbol{i}$ and $\boldsymbol{j}$ respectively. Since the nominal variables do not usually carry any inherent quantity, we suggest the following:

$$\begin{cases} \delta_{NOM}(A, A) = 0 \\ Otherwise, i.e. \delta_{NOM}(A, \neg A) = 1 \end{cases} \tag{4}$$

In other words, for the same nominal categories (e.g. two male respondents), the function evaluates to 0, and to 1 otherwise (male and female respondent). On the other hand, discrete ordinal variables, in their ordering capabilities, carry additional information that should be reflected in the distance metric, e.g. people of medium and high income are more similar than low and high income. In such a case, for $\boldsymbol{p}$ categories the variable is coded into, we define the following Toeplitz matrix (5), elements of which provide values for $\boldsymbol{\delta}$ for each combination of $\boldsymbol{g}^{th}$ and $\boldsymbol{h}^{th}$ ordinal categories, where $\boldsymbol{g}$ and $\boldsymbol{h}$ are row and column numbers respectively:

$$\begin{bmatrix} 0 & \frac{1}{p} & \frac{2}{p} & \dots & 1 \\ \frac{1}{p} & 0 & \frac{1}{p} & \dots & \frac{p-1}{p} \\ \frac{2}{p} & \frac{1}{p} & 0 & \dots & \frac{p-2}{p} \\ \vdots & \vdots & \vdots & \ddots & \frac{p-(r-1)}{p} \\ 1 & \frac{p-1}{p} & \frac{p-2}{p} & \frac{p-(c-1)}{p} & 0 \end{bmatrix} \tag{5}$$

Note that $\boldsymbol{r}$ and $\boldsymbol{c}$ are generic indices of row and column number respectively. The intuition behind (5) is that the difference between the most distant categories is valued the same as difference between any nominal variables while the intermediate values are proportional to their location relative to the extreme categories. Regardless of whether data is nominal or ordinal, if the compared value of a metric is a missing value, $\boldsymbol{\delta}$ function evaluates to 1 which is a conservative approach hedging against imputing missing values.

Using (4) and (5), along with marginal and joint probabilities of occurrence of particular categories, one can calculate the symbolic covariance $\boldsymbol{\Sigma_{Symb}}$ (McCane and Albert 2008). This, together with (3) can be used to obtain the symbolic covariance Mahalanobis distance $\boldsymbol{D}$:

$$D(\vec{x_i},\vec{x_j}) = \sqrt{\delta(\vec{x_i},\vec{x_j})^T \Sigma_{Symb}^{-1} \delta(\vec{x_i},\vec{x_j})} \quad (6)$$

Using (6) one can obtain distances between respondent $\boldsymbol{i}$ in the target dataset to each respondent $\boldsymbol{j}$ in the training dataset which determines $\boldsymbol{i}$'s set of $\boldsymbol{k}$ nearest neighbours. Various heuristics exist regarding the choice of $\boldsymbol{k}$ (see Robinson and Polak 2005; Hall et al. 2005) including adapting the value locally (Wang et al. 2005), yet arbitrary choice of a reasonable value of $\boldsymbol{k}$ is not uncommon, e.g. in traffic characterisation. In the context of this research, we have fixed the value of $\boldsymbol{k}$ to 15, yet the sensitivity of our final results to the choice of $\boldsymbol{k}$ remains subject of further research. Once the set of nearest neighbours is determined for the $\boldsymbol{i^{th}}$ respondent, a Monte Carlo draw is made to obtain missing variables with the probability of choosing a particular donor inversely proportional to its $\boldsymbol{D}$ (more similar respondents are also more likely to be donors). The number of draws is equal to the number of desired imputations, i.e. fused datasets. Clearly, generation of subsequent, fused datasets is a Markov process which simplifies the sensitivity analysis of the final results to the number of imputations. This is done in next section, and up to 5 imputations, which in the context of survey non-response would provide almost 95% efficiency (Rubin 1987). Once the datasets are obtained, the conventional analysis is performed and inferences drawn using (2). We present application of this fusion methodology to analysis of the interaction of ICT and physical mobility using structural equation modelling approach in the next section.

4. Application to Analysis of the Interaction of ICT and Physical Mobility

The data[1] used in the study comes from the US, UK and Norway and have been summarised in Table 1. The initial analysis, available in Pawlak et al. (2012), included a single dataset from Canada, but this part has been dropped as it did not deal with the issue of data fusion.

Table 1. Summary of the datasets used in the study

Country	Digital Behaviour (Year; Size)	Travel Behaviour (Year; Size)
US	PEW Internet Survey [a] (2007; 2 200)	American Time Use Survey [b] (2007; 12 248)
UK	ONS General Lifestyle Survey [c] (2010; 1 003)	National Travel Survey [d] (2010; 18 356)
Norway	ICT and Holiday Survey [e] (2005; 1 235)	Norwegian Travel Survey [f] (2005; 17 514)

[a]PEW 2007; [b]BLS 2007; [c]ONS 2010; [d]DfT 2010; [e]Statistics Norway 2005; [f]MMI Univero 2005.

[1] Some of the data applied in this publication are based on "National Travel Survey 2005, person file". The survey was financed by the Institute of Transport Economics. The data are provided by MMI Univero and prepared and made available by the Norwegian Social Science Data Services (NSD). Neither the Institute of Transport Economics, MMI Univero, nor NSD are responsible for the analysis/interpretation of the data presented here. Some of the data applied in this publication are based on "Travel and Holiday Survey, April 2005". The data are provided by Statistics Norway and prepared and made available by the Norwegian Social Science Data Services (NSD). Neither Statistics Norway, nor NSD are responsible for the analysis/interpretation of the data presented here. Analysis and interpretation of the other datasets are also entirely that of the authors.

The digital behaviour surveys are fairly diversified in terms of their character and content and usually smaller in scope to travel behaviour data, either time use or travel surveys. In the case of each country, there are a number of variables, mostly socio-demographic, which are shared by both digital and travel datasets (16 for the US, 17 for the UK, and 10 for Norway) and which have been shown in Table 2. A purpose-specific script for the fusion has been written in Ox (Doornik 2011), while the subsequent estimation of SEM has been done using LISREL (SSI 2012). The SEM approach has been chosen on the grounds of its flexibility in terms of handling numerous interrelated variables (Golob 2003), and its proven applicability in similar contexts as discussed in section 2.

Table 2. Shared variables used to calculate similarity measure

Country (No. of variables)	**Nominal**	**Ordinal**
US (16)	gender, marital status, state, urban area, race, Hispanic background, employment, student status, use of computer/Internet, use of e-mail, use of e-mail from home	age, education, presence/no. of children in household, household income,
UK (17)	gender, household type, type of tenure, employment status, region, self-employment status, supervisory role at work, ethnicity, use of bicycle/taxi/train/air travel, most frequent journey purpose	age, number of cars in the household, household income, frequency of bus use,
Norway (10)	gender, employment status, marital status, region	age, household size, work hours, household income, respondent income, education level

The models were evaluated using various fit indices, i.e. Chi-square, root mean square error of approximation (RMSEA) and its associated confidence interval. The results of our SEM estimation (cross-tabulated by country, relationship and number of imputations) are presented in Table 3. Given different levels of measurement and discrete coding of the variables, we opted for reporting the values as polychoric correlation coefficients (Jöreskog 1994, 2002) between various tele-activities ('T'activity') and frequency of travel for specific purposes ('Trav'purpose'). As the purpose of this paper is a methodological discussion, behavioural discussions is omitted, but see Pawlak et al. (2012) for a behavioural analysis in the context of Canadian data.

Clearly, the US results for single imputation suggest the existence of a number of significant relationships between digital lifestyle and travel behaviour such as complementarity effects between tele-working and work-related travel, tele-leisure and leisure related travel, substitution effects between tele-shopping and travel for shopping, or a negative correlation between tele-leisure and work-related travel (Table 3). However, as soon as the results are based on multiple datasets, half of the relationships become insignificant indicating the usefulness of the method in guarding against type 1 error. A peculiar case of the relationship between tele-working and leisure-related travel: the initially significant result becomes insignificant in case of using 2 and 3 datasets, but becomes significant again when 4 and 5 datasets are used. The reason for that might be two-fold: firstly the value of the correlation coefficient itself is quite small, and secondly its standard error is quite high. As a result, the results appear to oscillate around the significance cut-point. Assuming conservative approach,

Table 3. SEM estimation results and their sensitivity to number of imputations used ***m***

	US					UK					Norway				
m	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
T'wrk[a]/ Trav'wrk	0.10	0.12	0.10	0.10	0.10	*0.01*	*0.03*	*0.02*	*0.02*	*0.03*	0.19	0.21	0.22	0.21	0.21
T'shop[b]/ Trav'shop	-0.16	*-0.12*	*-0.07*	*-0.07*	*-0.07*	*0.08*	*0.09*	*0.09*	*0.08*	*0.04*	.	.	.	.	.
T'leis[c]/ Trav'leis	0.25	0.25	0.26	0.28	0.29	*0.05*	*0.02*	*0.01*	*-0.01*	*-0.03*	*0.03*	*0.03*	*0.03*	*0.03*	*0.03*
T'leis/ Trav'escrt[d]	.	.	.	.	.	.	.	.	.	.	0.14	0.14	*0.11*	*0.11*	*0.11*
T'conf[e]/ Trav'conf	-0.11	*-0.06*	*-0.10*	*-0.10*	*-0.11*	.	.	.	.	.	.	.	.	.	.
T'soc[f]/ Trav'soc	.	.	.	.	.	*-0.02*	*-0.09*	*-0.07*	*-0.07*	*-0.08*	0.16	*0.07*	*0.08*	*0.05*	*0.04*
T'serv[g]/ Trav'serv	.	.	.	.	.	.	.	.	.	.	*-0.02*	*-0.05*	*-0.07*	*-0.05*	*-0.04*
Blog crt[h]/ Trav'soc	-0.12	*-0.04*	*-0.04*	*0.00*	*0.03*	.	.	.	.	.	.	.	.	.	.
Blog rd.[i]/ Trav'soc	0.09	*0.05*	*0.03*	*0.00*	*-0.02*	.	.	.	.	.	.	.	.	.	.
T'wrk/ Trav'shop	0.12	0.10	*0.06*	*0.07*	*0.07*	.	.	.	.	.	0.15	*0.11*	*0.09*	*0.08*	*0.07*
T'wrk/ Trav'leis	0.07	*0.07*	*0.09*	0.10	0.11	.	.	.	.	.	.	.	.	.	.
T'leis/ Trav'wrk	-0.37	-0.37	-0.39	-0.37	-0.37	.	.	.	.	.	.	.	.	.	.
T'leis/ Trav'shop	-0.20	-0.21	-0.19	-0.19	-0.20	.	.	.	.	.	.	.	.	.	.
T'leis/ Trav'conf	-0.21	-0.20	-0.19	-0.19	-0.20	.	.	.	.	.	.	.	.	.	.
T'leis/ Trav'soc	-0.16	-0.12	-0.13	-0.15	-0.16	.	.	.	.	.	.	.	.	.	.
T'conf/ Trav'leis	.	.	.	.	.	-0.17	*-0.07*	*-0.05*	*-0.03*	*-0.02*	.	.	.	.	.
T'shop/ Trav'oth[j]	.	.	.	.	.	-0.23	*-0.16*	*-0.13*	*-0.11*	*-0.10*	.	.	.	.	.
T'conf/ Trav'shop	.	.	.	.	.	.	.	.	.	.	-0.17	*0.01*	*-0.11*	*-0.10*	*-0.09*
T'soc./ Trav'serv	.	.	.	.	.	.	.	.	.	.	-0.21	*-0.14*	*-0.14*	*-0.14*	*-0.14*
SatNav[k]/ Trav'serv	.	.	.	.	.	.	.	.	.	.	0.20	*0.09*	*0.06*	*0.03*	*0.01*

Notes: Values shaded and in italics are significant at 95% level. The empty cells indicate that either the relationship could not be investigated due to lack of suitable variables, or the estimated correlation was insignificant and close to zero. **Key:** [a]work; [b]shopping; [c]leisure; [d]escorting; [e]conferencing; [f]socialising; [g]services; [h]creation; [i]reading; [j]other; [k]posession of a satellite navigation set.

such an estimate should not be considered robust. In the case of the UK, only two relationships and only for a single imputation appear significant. However, it remains challenging to determine the source of insignificance, i.e. does it result from imperfections in the data fusion method, or simply from non-existence of the relationships? In fact, while guarding against type 1 error, the method might be prone to higher risk of type 2 errors. In the case of Norway, one can identify a number of significant relationships for a single

imputation, but only the strong and stable complementarity effect between tele-working and travel for work purpose remain significant if multiple imputations framework is applied.

There are, however, limitations of the approach, and these are clearly demonstrated in Figure 1 which relates the percentage increase in the width of confidence intervals to the percentage of missing information, i.e. ratio between between-imputation variance (4) to the total variance (5). Strikingly, the results for all the countries analysed follow a similar, logarithmic path. One can see that until the percentage of missing information is somewhat below 50%,

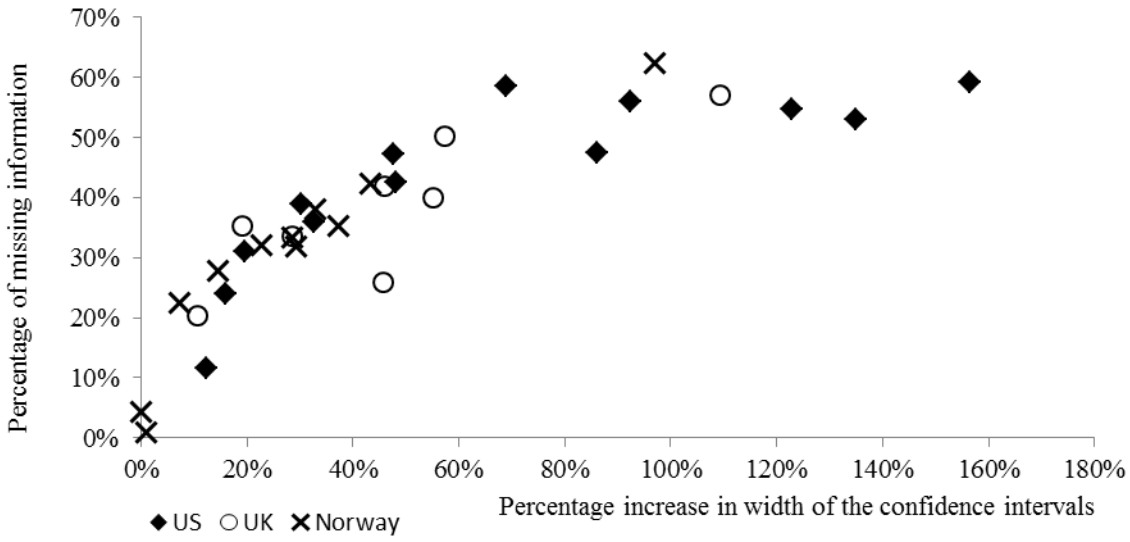

Figure 1. Relationship between the size of confidence intervals and missing information

the expected increase in confidence intervals remains approximately below 60%. Above that threshold, the widths of the intervals soar and, unless the analysed relationship is really strong and the shared variables accurately propagate it between the datasets, it is unlikely to be captured in the fused datasets. A possible solution would be to include more shared variables and investigate their explanatory power in the imputed variables to determine a priori to what extent this fusion technique is applicable. The former in itself does not guarantee better results (note that the UK has more shared variables than the US, yet no significant estimate in the final analysis) due to the possibility of the curse of dimensionality. This fairly common issue in the k-NN approach results from the reduction in the density of the feature space as number of dimensions rises for a fixed number of observations (Aluja-Banet et al. 2007). The effect can be further enhanced when the shared variables are not weighted appropriately to reflect their relevance. Noting this vulnerability of the k-NN approach, Aluja-Banet et al. (2007) suggest reduction of the dimensionality of the vector of common variables using, e.g. factorial analysis which also decreases the cost of the implementation. On the other hand, we suggest a method, still being developed and tested, based on the extreme value theory (EVT), which could not only provide the necessary weights for the variables (arguably guarding against the aforementioned difficulty), but also be a convenient way of testing to what extent fusion of particular datasets is feasible.

Let us consider the donor dataset only. We suggest that the usefulness of the method can be deduced by considering the probability of accurate identification of an individual $\boldsymbol{f}$ within the training dataset, knowing only his or her values of the shared variables $\boldsymbol{X_{SHA}}$. If this

probability is high, it means that the shared variables can identify the person accurately and the method described above can credibly impute the variables from the training dataset. Let us further assume that in order to be accurately identified, the value of the distance metric $\boldsymbol{D}$ based on $\boldsymbol{X_{SHA}}$ needs to be smaller than for any other individual $\boldsymbol{j}$ in the training dataset:

$$P\left(D_f\left(X_{fSHA}\right) + \varepsilon_f < D_j\left(X_{jSHA}\right) + \varepsilon_j \right) \quad \forall j \tag{7}$$

The error term $\boldsymbol{\varepsilon}$ reflects the fact that $\boldsymbol{X_{SHA}}$ might not capture all the relevant features necessary to identify an individual. In fact, (7) is an expression linking the problem to the extreme value theory. Given appropriate distributional assumptions, it can be used to formulate a model, similar to the well-established discrete choice models (Ben-Akiva and Lerman 1985), and estimates the relative relevance (weights) of the shared variables in identifying an individual in the whole training dataset, in a similar manner to the validation performed by Saporta (2002). However, by introducing additional parameter $\boldsymbol{\theta}$ in (7):

$$P\left(D_f\left(X_{fSHA}\right) + \varepsilon_f < D_j\left(X_{jSHA}\right) + \theta + \varepsilon_j \right) \quad \forall j \tag{8}$$

where the $\boldsymbol{\theta}$ is an assumed distance defining the neighbourhood (locality) of the individual $\boldsymbol{f}$. As such, (8) would provide the way of calculating the probability that a particular neighbour $\boldsymbol{j}$ lies within the locality of $\boldsymbol{j}$. Since the k-NN approach relies on the assumption that the values are drawn locally (Aluja-Banet 2007), (8) provides additional way of investigating the robustness of the fusion method suggested above using well-established EVT-based tools.

In fact, the problem is similar in nature to estimating parameters describing relative importance of attributes of discrete alternatives, such as transport modes. Clearly, the sheer size of the training (choice) set would require reduction (pruning) methods which could also provide guidance in determining the optimal number of $\boldsymbol{k}$ nearest neighbours. However, fit-indices, e.g. rho-squared, could guide optimal combination and relative weighting of the variables in the distance metric on one hand and indicate their explanatory power, thus allowing the researcher to evaluate the usefulness of the implicit fusion. While the method is still developed and tested, we perceive it as a promising direction, offering chance to re-use the existing datasets in novel ways.

5. Conclusions

Our analysis above indicates that a k-NN based multiple imputations approach to the implicit fusion of data might offer a promising way of re-using the existing data by combining datasets whose collection in a single survey might either be costly, impossible due to the historical nature of the data, or other constraints. Using datasets from 3 countries, the US, UK, and Norway in the context of investigating the relationships between digital and travel behaviour, we demonstrate that the method is capable of capturing some apparently existing relationships while fairly robustly guarding against type 1 error as compared to analysing a single fused dataset. In doing so, however, the method appears more prone to type 2 errors. As a final contribution, we present possible improvements to the technique based on the extreme value theory which, although still developed and tested, can possibly open even more possibilities to efficiently re-use existing data sources in new, innovative ways.

References

Aluja-Banet, T., Daunis-i-Estadella, J., and Pellicer, D. (2007). GRAFT, Complete System for Data Fusion. Computational Statistics & Data Analysis, 52, 635-649.

Andreev, P., Salomon, I. and Pliskin, N. (2010). Review: State of Teleactivities. Transportation Research C: Emerging Technologies, 18, 3-20.

Axhausen, K.W. (1998). Can We Ever Obtain the Data We Would Like to Have?. In Theoretical Foundations of Travel Choice Modeling, (eds). T. Gärling, T. Laitila and K. Westin, Amsterdam: Elsevier.

Axhausen, K.W. (2008). Social Networks, Mobility Biographies, and Travel Survey Challenges. Environment and Planning B: Planning and Design, 35, 981-996.

Ben-Akiva, M. and Lerman, S.R. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. London: MIT Press.

BLS (2007). American Time Use Survey 2007. Dataset. Washington D.C.: Bureau of Labor Statisitcs. Available at http://www.bls.gov/tus/. (Accessed February 2012.)

Choo, S., Lee, T., and Mokhtarian, P.L. (2008). Do Transportation and Communications Tend to Be Substitutes, Complements or Neither? U.S. Consumer Expenditures Perspective, 1984-2002. Transportation Research Record, 2010, 121-132.

D'Ambrosio, A., Aria, M., and Siciliano, R. (2007) Robust Tree-Based Incremental Imputation Method for Data Fusion. Lecture Notes in Computer Science, 4723, 174-183.

DfT (2010). United Kingdom National Travel Survey. Dataset. London: Department for Transport.

de Leon, A.R. and Carrière, K.C. (2005). A Generalized Mahalanobis Distance for Mixed Data. Journal of Multivariate Analysis, 92, 74-185.

Doornik, J.A. (2011) OxEdit. Software. Available at http://www.doornik.com/ox. (Accessed September 2011).

Golob, T.F. (2003). Structural Equation Modeling for Travel Behaviour Research. Transportation Research Part B: Methodological, 37, 1, 1-25.

Golob, T.F. and Regan, A.C. (2001). Impacts of Information Technology on Personal Travel and Commercial Vehicle Operations: Research Challenges and Opportunities. Transportation Research C: Emerging Technologies, 9, 87-121.

Hall, P., Park, B.U., and Samworth, R.J. (2008). Choice of Neighbor Order in Nearest-Neighbor Classification. The Annals of Statistics, 36, 5, 2135-2152.

Hamermesh, D.S., Frazis, H., and Stewart, J. (2005). Data Watch: American Time Use Survey. Journal of Economic Perspectives, 19, 1, 221-232.

Jöreskog, K.G. (1994). On the Estimation of Polychoric Correlations and their Asymptotic Covariance Matrix. Psychometrika, 59, 3, 381-389.

Jöreskog, K.G. (2002). Structural Equation Modeling with Ordinal Variables Using LISREL. Skokie, IL: SSI. Available at: http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf. (Accessed May 2012.)

Lyons, G. and Urry, J. (2005). Travel Time Use in the Information Age. Transportation Research Part A: Policy and Practice, 39, 257-276.

McCane, B. and Albert, M. (2008). Distance Functions for Categorical and Mixed Variables. Pattern Recognition Letters, 29, 986-993.

MMI Univero (2005). Norwegian National Travel Survey, person file. Dataset.

Mokhtarian, P.L. (1990). A Typology of the Relationships between Telecommunications and Transportation. Transportation Research Part A: General, 24a, 231-242.

Noce, A.A. and McKeown, L. (2008). A New Benchmark for Internet Use: A Logistic Modelling of Factors Influencing Internet Use in Canada, 2005. Government Information Quarterly, 25, 462-476.

ONS (2010) ONS Opinions and Lifestyle Survey, January-April 2010. Dataset. London: Office for National Statistics.

Pawlak, J., Polak, J.W., and Sivakumar, A. (2011). The Consequences of the Productive Use of Travel Time: Revisiting the Goods-Leisure Trade-off in the Era of Pervasive ICT. Paper prepared for the International Choice Modelling Conference, 4-6 July, Leeds, United Kingdom.

Pawlak, J., Sivakumar, A., and Polak, J.W. (2012). Digital Behaviour and Physical Mobility: A Cross-country Structural Equation Approach. Paper prepared for the International Association for Travel Behaviour Conference, 15-20 July 2012, Toronto, Canada.

PEW (2007). February-March 2007 Tracking. Dataset. PEW Research Center. Available at http://pewinternet.org/Shared-Content/Data-Sets/2007/FebruaryMarch-2007-Tracking.aspx. (Accessed in Feburary 2012.)

Robinson, S. and Polak, J.W. (2005). Modelling Urban Link Travel Time with Inductive Loop Detector data using the k-NN method. Transportation Research Record, 1935, 47-56.

Rubin, D.B. (1987). Multiple Imputation for Non-response in Surveys. Chichester: John Wiley & Sons.

Salomon, I. (1986). Telecommunications and Travel Relationships: A Review. Transportation Research Part A: General, 20, 3, 223-238.

Salomon, I. (2000). Can Telecommunications Help Solve Transportation Problems? In Handbook of Transport Modelling, (eds). D.A. Hensher and K. Button, Oxford: Pergamon Press.

Saporta, G. (2002). Data Fusion and Data Grafting. Computational Statistics & Data Analysis, 38, 465-473.

Senbil, M. (2009). Forecasting Information and Telecommunications Technologies in an Era of Change: What Are the Implications for Travel? Middle East Technical University Journal of the Faculty of Architecture. 26, 2, 139-152.

Singh, P., Paleti, R., Jenkins, S., and Bhat, C.R. (2011). On Modeling Telecommuting Behavior: Option, Choice, and Frequency. Paper prepared for the 91st Annual Meeting of the Transportation Research Board, 22-26 January 2012, Washington D.C., USA.

Sivakumar, A. and Polak, J.W. (2013). 'Exploration of Data-Pooling Techniques: Modeling Activity Participation and Household Technology Holdings'. Paper prepared for the 92nd Annual Meeting of the Transportation Research Board, 13-17 January 2012, Washington D.C., USA.

SSI (2012) LISREL Student Edition. Software. Skokie, IL: SSI. Available at: http://www.ssicentral.com/lisrel. (Accessed January 2012.)

Statistics Norway (2005). Travel and Holiday Survey, April 2005. Dataset.

Stopher, P.R. (1992). Use of an Activity-based Diary to Collect Household Travel Data. Transportation, 19, 159-176.

Stopher, P.R. (2000). Survey and Sampling Strategies. In Handbook of Transport Modelling, (eds). D.A. Hensher and K. Button, Oxford: Pergamon Press.

Wang, J., Neskovic, P. and Cooper, L.N. (2005). Locally Determining the Number of Neighbors in the k-Nearest Neighbor Rule Based on Statistical Confidence. Lecture Notes in Computer Science, 36190, 71-80.

Wang, D. and Law, F. (2007). Impacts of Information and Communication Technologies (ICT) On Time Use and Travel Behaviour: A Structural Equation Analysis. Transportation, 34, 513-527.